Forecasting Development

Forecasting Research Technical Report No. 379

Spatial and lead–time accuracies of thunderstorm forecasts for air traffic control

April 2002

H.A. Watkin, T.R. Scott and D.J. Hoad

© Crown Copyright 2002

Met Office , Forecasting Development , Aviation Technology Centre, Room R321 , London Road , Bracknell , Berkshire ,RG12 2SZ, United Kingdom

00/0486

Forecasting Research
Technical Report No. 379

Spatial and lead-time accuracies of thunderstorm forecasts for air traffic control

by

H.A. Watkin, T.R. Scott and D.J. Hoad

April 2002

The Met Office
Forecasting Development
Aviation Technology Centre
Room R321
London Road
Bracknell
RG12 2SZ
United Kingdom

© Crown Copyright 2002

Permission to quote from this paper should be obtained from the above Met Office division

Please notify us if you change your address or no longer wish to receive these publications.

Tel: 44 (0)1344 856245 Fax: 44 (0)1344 854026 email: jsarmstrong@meto.gov.uk

SPATIAL AND LEAD-TIME ACCURACIES OF THUNDERSTORM FORECASTS FOR AIR TRAFFIC CONTROL

Helen A. Watkin, Tania R. Scott, and Deborah J. Hoad

April 2002

ABSTRACT

This project aims to quantify the trade-off between the spatial and lead-time accuracies of forecasts of air traffic disruption due to thunderstorm activity, for three different types of Met Office forecasts. The forecasts are tested for 10 thunderstorm events that occurred over the UK during summer 2001. Forecast data are used to replicate what a pilot would see on their radar display, as the coverage and intensity of reflectivities on these displays dictates the decision as to whether or not to deviate around a particular area. Levels of air traffic disruption are classified depending on the percentage, of the area of a given air space, predicted to cause deviations. Verification is carried out against disruption levels calculated from radar observations, and forecast performance assessed at each combination of four spatial resolutions and hourly lead times up to T+12. Forecast quality is shown to be greatest for disruption forecasts prepared from Nimrod data and following a method which uses both storm intensity and coverage to predict deviations. Forecast accuracy and skill decline with increasing lead time and spatial resolution up to T+5, and remain fairly constant at lead times greater than this. Forecast quality is shown to be dependent on the synoptic conditions, and the results affected by limitations in the forecasting and the verification methods.

1 INTRODUCTION

For reasons of safety and passenger comfort aircraft often fly around thunderstorms, and thus air traffic management (ATM) need advance notice of thunderstorm activity to enable them to adjust air traffic flow before major disruption occurs. This project assesses the Met Office's ability to forecast air traffic disruption due to thunderstorm activity.

A previous Met Office project investigating short range thunderstorm forecasts from the GANDOLF (Generating Nowcasts for the Deployment of Operational Land-based Flood forecasts) system, showed that GANDOLF forecasts could provide useful information up to about 90 minutes ahead (Hoad, 2000). Feedback on the project from the UK ATM authority suggested they would like more information on the quality of forecasts at a number of different spatial and temporal scales. Hence, this follow-up project aims to quantify the trade-off between the spatial and lead-time accuracies of Met Office disruption forecasts. The spatial resolutions at which the disruption forecasts are prepared are 15km, 30km, 60km and 120km. The forecast lead times start with an analysis (T+0), then proceed hourly up to T+12.

Air traffic disruption occurs when pilots perceive a threatening storm ahead and decide to deviate around it. Studies of pilots' storm penetration/deviation behaviour have shown storm intensity and coverage to be two key decision factors (Rhoda et al, 2000; Rhoda and Pawlak, 1999). Information on these factors is largely taken from aircraft radar reflectivities, which are displayed in the cockpit using three colours: green (corresponding to a rain rate of 1-4mm/hr), amber (4-12mm/hr) and red (>12mm/hr) (Seymour, 2002). Using this display a pilot will look at the maximum intensity of an upcoming storm, both horizontally and vertically, to determine if it is dangerous. UK pilots will not fly through amber or red and also try to avoid green where possible, but will fly through green if it is difficult to avoid the area (Rankin, 2001), confirming the importance of both intensity and coverage.

To create accurate forecasts of the level of disruption, this paper attempts to take account of pilot behaviour in its forecasting methods. The forecast and verification data are used to replicate pilot radar displays, which are used to allocate the penetration/deviation behaviour. Levels of air traffic disruption are classified depending on the percentage of the area of a given air space that is shown to cause deviations. The relative performance of the disruption forecasts produced using different methods, forecast data types, spatial resolutions and lead times are discussed, to allow ATM authorities to determine the value and utility of the forecasts in the context of their own operations.

Section 2 of this report describes the forecasting and verification data used and the methods followed to produce and verify the forecasts of air traffic disruption due to thunderstorm activity. An analysis of forecast quality measures is contained in section 3. Section 4 discusses the trends in forecast quality and the limitations of the project, with the conclusions of the report being presented in section 5.

2 METHOD

2.1 Identifying thunderstorm events

Figure 1 shows how lightning data (shown by crosses) were used to identify thunderstorm events across the UK during summer 2001. The locations and range of the radar sites used in the verification of this project are also shown. For each event, air traffic disruption was forecast for an area of 240km^2 centred on the radar site that best tracked the storm progress, in this case Hameldon Hill. Suitable archived data were not available for all of the selected events, but 10 events were suitable for analysis, and are listed in appendix A.

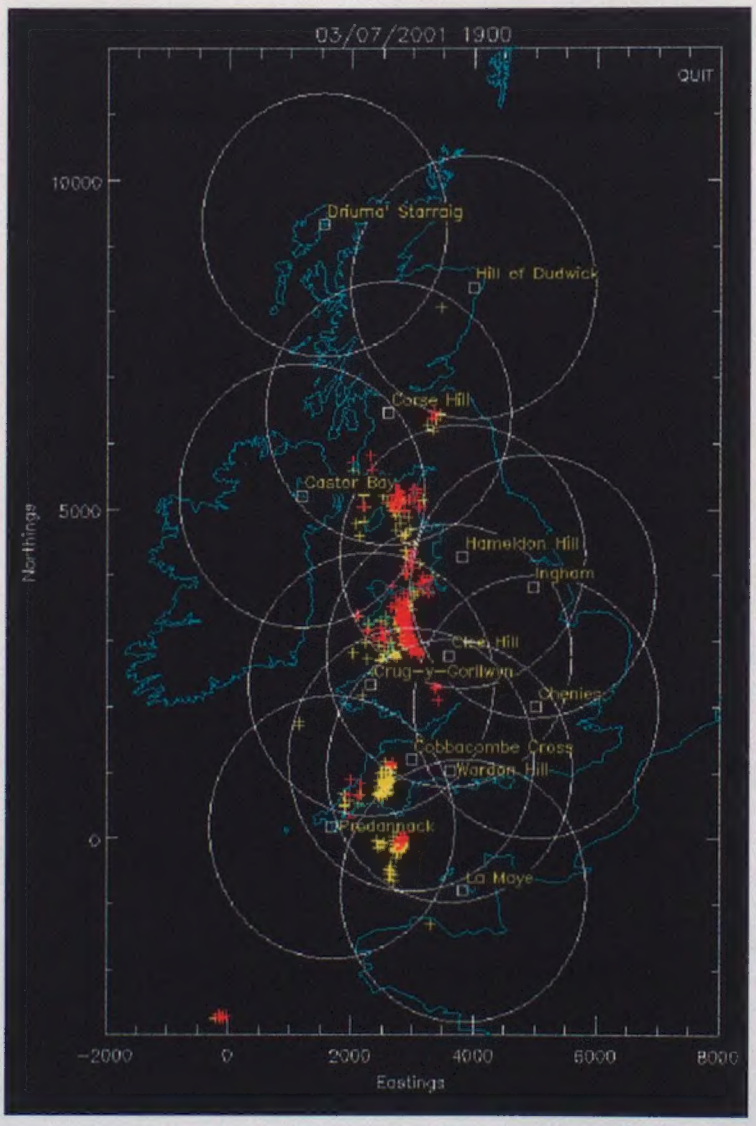


FIGURE 1: Image indicating the method used to identify thunderstorm events for analysis. The crosses are lightning strikes located by the Met Office's lightning detection system during a one hour period. The colour of each cross indicates the status of the system, which can be overloaded by a high strike rate so that strikes are missed. Yellow, orange, and red are for lightning strikes that were observed when none, some, and many strikes were missed.

2.2 Replicating pilot radar displays

Pilot radar displays were replicated using both forecast and radar verification data. This was done by estimating the maximum rain rate in the cloud, as there is a quantifiable relationship between radar reflectivity and rain rate. A 15km resolution was chosen to match the resolution of several of the forecast data fields and to correspond to the size of large thunderstorms in the UK, which can be between 10 and 20km wide.

Three types of forecast data were used to predict the convective rain rate at the ground:
i. UK Mesoscale Model The Mesoscale numerical weather prediction model is a limited-area model of the Met Office Unified Model. It covers the British Isles and surrounding areas with a spatial resolution of 12km, and is run every 6 hours. Mesoscale convective rain rate forecast data were converted to a 15km resolution.
ii. Nimrod nowcasting system The Nimrod nowcasting system integrates recent observations with Mesoscale Model products to provide frequent forecasts over the UK up to 6 hours ahead (Golding, 1998). In this case a forecast of total rain rate at the ground was used because convective rain rate was not available. Dynamic rain is eliminated by masking with convective cloud cover from the Mesoscale model.
iii. CDP The Convection Diagnosis Project (CDP) provides forecast probability diagnoses of the intensity, distribution and duration of convective showers from Mesoscale model outputs (Hand, in submission). It runs every 6 hours providing hourly forecast products from T+6 hours. The forecast data used were for peak rainfall rate for probabilities of shower occurrence of greater than 70%.

These values were then converted into the maximum rain rate in the cloud using the idealised rain rate profile of a mature thunderstorm cell shown in figure 2. The profile was scaled logarithmically from the average figures found in UK thunderstorms of 6mm/hr at the ground corresponding to a 16mm/hr maximum in the cloud (Hand, 1996).

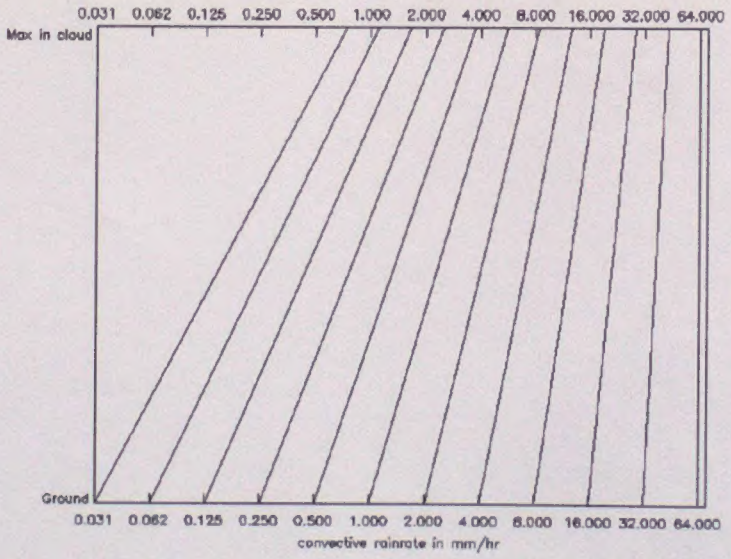


FIGURE 2: Idealised vertical rain rate profiles for mature cells, between the ground and the location of maximum rain rate in the cloud.

Radar verification data
The radar data used for the verification were taken from the nearest radar site to the event in question, and from any suitable neighbouring sites. Data from four beams were available, each at a different elevation angle and hence reflecting from a different height in the storm. For a radar to measure the rain rate in a convective cloud at least one of the radar beams must be above the convective cloud base. From Mesoscale forecasts of convective rain rates and their corresponding convective cloud bases, it was found that convective clouds generally have cloud bases at about 3km or below, especially those with high rain rates. Thus radar data used for forecast verification should

be for beam heights of at least 3km. These beam heights only occur well away from a radar site, limiting the coverage of any given radar, and hence the amount of data available for forecast verification. However the overall coverage can sometimes be improved by combining data from several neighbouring radar sites. The height of all of the available beams at each pixel location was calculated and discarded if it was below 3km or above 15km. The maximum reflectivity from all of the suitable beams was then taken and converted into mm/hr to give an estimate for the maximum rain rate in the cloud. If no beam could be found at a suitable height this pixel was masked and not included in the analysis.

Classifying by radar colour

The maximum rain rates for the three forecasts, and the radar data were then classified into the colours they would appear on the aircraft radar. A total area of 240 by 240 km (16 by 16 grid cells) was analysed for each event, centred on the radar site closest to the storm activity. Figure 3 shows an example of replicated pilot radar displays over Eastern England.

FIGURE 3: Replicated pilot displays for 1500 on 15 June 2001, created using Nimrod forecast data at T+1 (left), and radar data (right). A 240km^2 area centred on the Ingham radar site is shown. Black corresponds to a rain rate under 1mm/hr, green to 1-4mm/hr, amber to 4-12mm/hr, and red to greater than 12mm/hr.

2.3 Calculating air traffic disruption

Once replicated pilot displays were obtained, the data were analysed for air traffic disruption. This was done in the same way for the three forecast types (Mesoscale, Nimrod, and CDP) and for the radar data, to ensure a fair and consistent verification procedure. Two methods for calculating air traffic disruption were used, one taking into account storm intensity, and the other taking into account intensity and coverage.

2.3.1 Intensity method

Rain rate pixels were classed as deviation pixels if their aircraft radar classification was amber or red (i.e. above 4mm/hr). The level of air traffic disruption was categorised for the four spatial resolutions (covered by 1, 4, 16, and 64, 15km pixels as shown in figure 4) according to the percentages of deviation pixels in that region. Table 1 gives the definition of each level of disruption.

% of deviation pixels (D)	Level of disruption
D=0	None
0<D<25	Minor
25≤D<37.5	Major
D≥37.5	Severe

TABLE 1: Classifying the level of air traffic disruption based on the percentage of the area classified as causing a pilot to deviate around it, based on the intensity method.

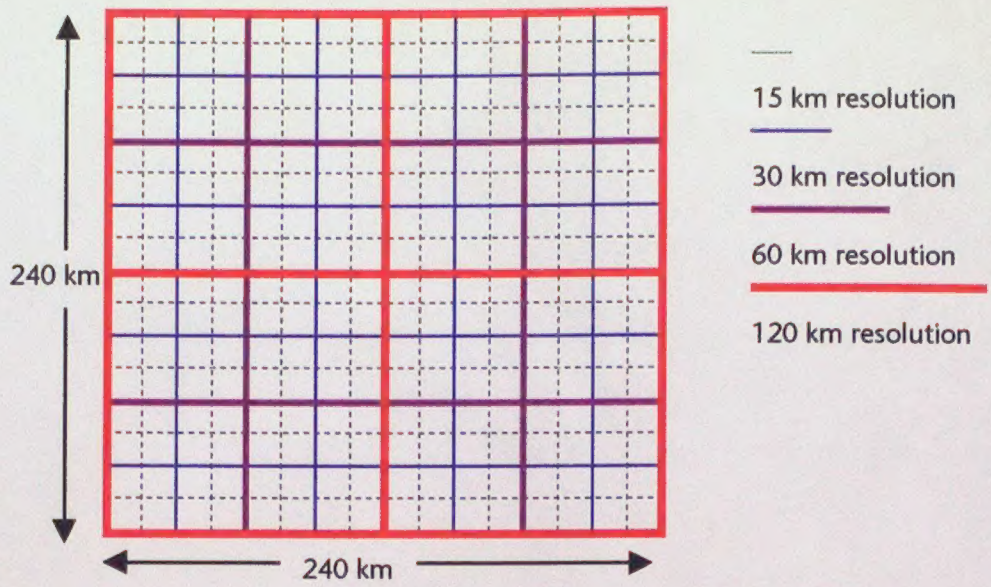


FIGURE 4: Diagram showing the four spatial resolutions for which air traffic disruption forecasts are produced, plotted on the 240km^2 analysis area used in each event.

2.3.2 Intensity and coverage method

The type of convective activity in the 75 by 75 km area surrounding each 15km pixel was classified as being not widespread, widespread and weak, or widespread and strong according to the criteria given in table 2. The time taken to fly across an area of this size is approximately 5 minutes, assuming a cruise speed of about 250 m/s, allowing a reasonable timescale for decision making. Table 2 shows how each pixel was classified as leading to "no deviation (N)", "small deviation (S)", or "large deviation (L)" from the expected flight path, depending on its combination of on-board radar display colour and coverage type.

Radar return	Type of convective activity in nearby area		
	Not widespread <20% non-zero pixels	Widespread weak >20% non-zero, <20% amber-red	Widespread strong >20% non-zero, >20% amber-red
Black	N	N	N
Green	S	N	L
Amber-Red	S	S	L

TABLE 2: Estimating grid pixel deviation characteristics in the intensity and coverage method. No deviation (N) was allocated if the pixel radar return was black, independent of the type of convective activity. Pilots avoid green radar returns if possible (thus a small deviation (S) was allocated when coverage is not widespread), but fly through it if it is difficult to avoid the area (causing no deviation in widespread weak convection). If the convective activity is widespread and strong the pilot will probably avoid the whole area causing a large deviation (L). Amber or red radar returns are all classified as causing a deviation, with this deviation being small unless there is widespread strong activity when a large deviation is allocated.

Table 3 shows how the level of air traffic disruption in each airspace was classified into the same four categories as in the simpler method, but this time using percentages of the deviation types classified in table 2.

%Large (L)	Operator	%Total (T) NB. T=S+L	Disruption
L=0	AND	T=0	None
L≤12.5	AND	T<25	Minor
L<25	AND	T<37.5	Major
L≥25	OR	T≥37.5	Severe

TABLE 3: Classifying the level of air traffic disruption based on the percentage of the different deviation types estimated in the intensity and coverage method.

Contingency tables were created from the results, for the different forecasts, methods, lead times and spatial resolutions. These tables are 4 x 4, because there are four disruption classifications.

3 ANALYSIS OF THE RESULTS

To analyse the results several statistical forecast quality measures were calculated from the contingency tables using the methods and formulae defined in Wilks (1995). Appendix B gives an example contingency table and the formulae, while table 4 summarises the measures. The Hit Rate and the Kuipers Skill Score give one value for each 4 x 4 contingency table whereas the other measures give a separate value for each of the disruption classes.

Measure	Definition	Numbers
Hit Rate	fraction of forecasts correct	perfect forecast=1 completely wrong=0
Critical Success Index (CSI)	hit rate once 'disruption type not forecast and not observed' values are removed	perfect forecast=1 completely wrong=0
Probability of detection (POD)	fraction of times disruption type was forecast when it had been observed	perfect forecast=1 completely wrong=0
False Alarm Rate (FAR)	fraction of times disruption type was forecast to occur and didn't occur	perfect forecast=0 completely wrong=1
Bias Ratio	the ratio of the number of times a disruption category was forecast to the number of times it was observed	over-forecast if >1 under-forecast if < 1
Kuipers' Skill Score (KSS)	a skill score calculated by comparing the accuracy of the forecasts with unbiased random forecasts	perfect forecast=1 random forecasts=0 forecast inferior to random if < 0

TABLE 4: Definitions of the statistical forecast quality measures used in this project.

3.1 Comparison of the two disruption classification methods

The two disruption classification methods (intensity only vs intensity and coverage) perform similarly in terms of Hit Rate and KSS, but figure 5 shows that the method which incorporates both intensity and coverage has a smaller degree of bias, with all categories lying within ±0.9 of the unbiased value of 1. The 'Severe' category is still over-forecast, but to a more acceptable level. Figure 6 shows the CSI and POD scores, averaged over spatial resolution, for the forecasts produced from Nimrod data in the 'Severe' category, which are particularly important in terms of ATM impact. The CSI scores for the method including intensity and coverage are raised by an average of around 0.2, and the drop off in POD scores with increased lead time is slower in the more complex method. Thus the forecasts produced using both intensity threshold and coverage levels are of the greater quality. The remaining analysis will therefore be for the more complex method only.

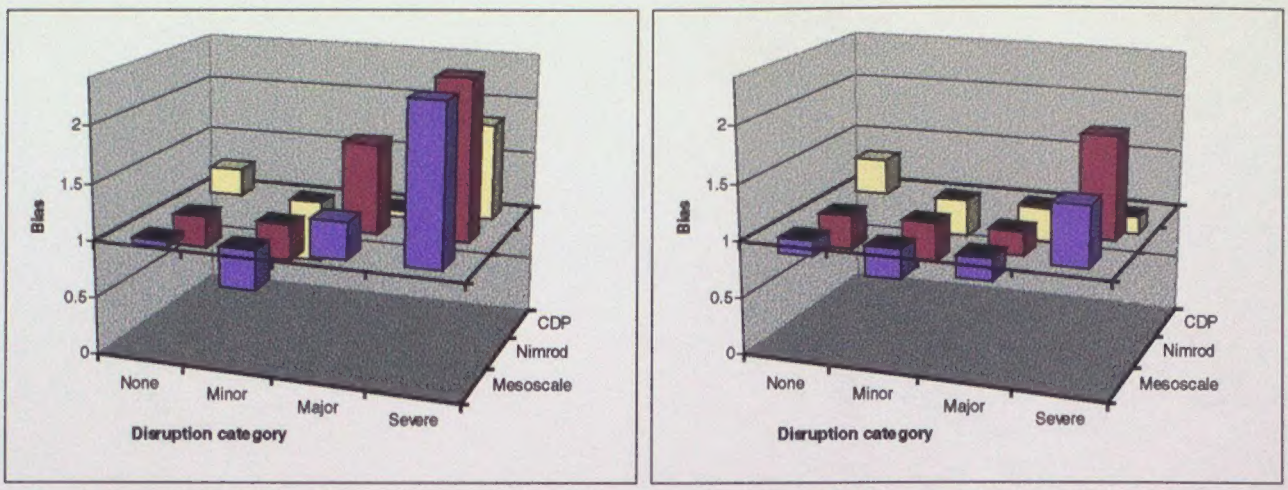


FIGURE 5: Bias values (averaged over lead time and resolution) for each forecast type and disruption category for the intensity-only method (left) and the intensity and coverage method (right). The severe category is over-forecast in both methods, but to a lesser degree in the intensity and coverage method.

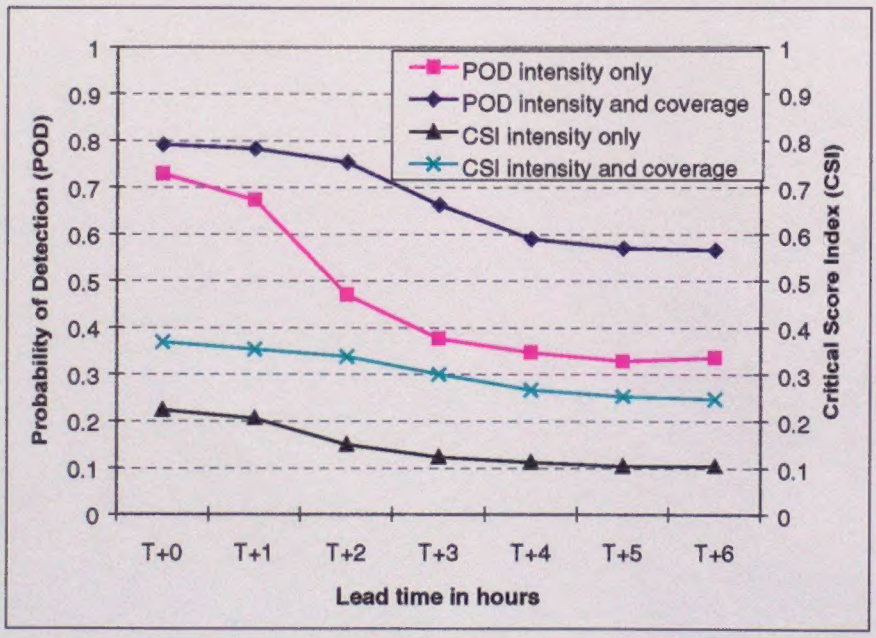


FIGURE 6: POD and CSI results for the 'Severe' class of forecast prepared from Nimrod data using the two different methods. The absolute level of both measures is higher for the intensity and coverage method, with this method also showing a slower decline in the POD scores with increased lead time.

3.2 Comparison of the different forecast types

Disruption forecasts generated using CDP data vary greatly in their quality between cases, with some of the thunderstorm events not being forecast at all, while others are severely over-forecast. Thus only disruption forecasts prepared from Mesoscale and Nimrod data will be compared.

The difference in forecast quality between these two forecast types depends on the quality measure. Figure 7 shows the average Hit Rate and Kuipers' Skill score for both forecasts, with best fit trend lines fitted. The Hit Rate for each lead time varies between 0.6 and 0.45 for both forecasts. Both trend

lines show a decline over lead time, with a steeper drop off for the Nimrod forecasts. The Kuipers Skill Scores show significant differences. Both forecasts decline with lead time before seeming to level out at around 0.1. However, the Nimrod forecasts perform significantly better than the Mesoscale data up to T+3. The Nimrod best fit line is a cubic regression, which adequately follows the pattern of the data, declining slowly with lead time at first, then more steeply, before levelling out again around T+6. This pattern is probably a consequence of the weighting between observations and Mesoscale data in the Nimrod forecasts, with the positive input of observations having less influence over longer lead times. The scatter in the Mesoscale forecasts makes them harder to interpret, but a best fit linear regression line shows a clear drop off with increased lead time.

The scatter in the Mesoscale forecasts is due to there being a smaller sample of data for this forecast type than for the Nimrod forecasts. The reasons for this disparity are discussed in Section 4.1 along with the consequences for the Mesoscale forecasts of having a greater sensitivity to the individual event and time in question.

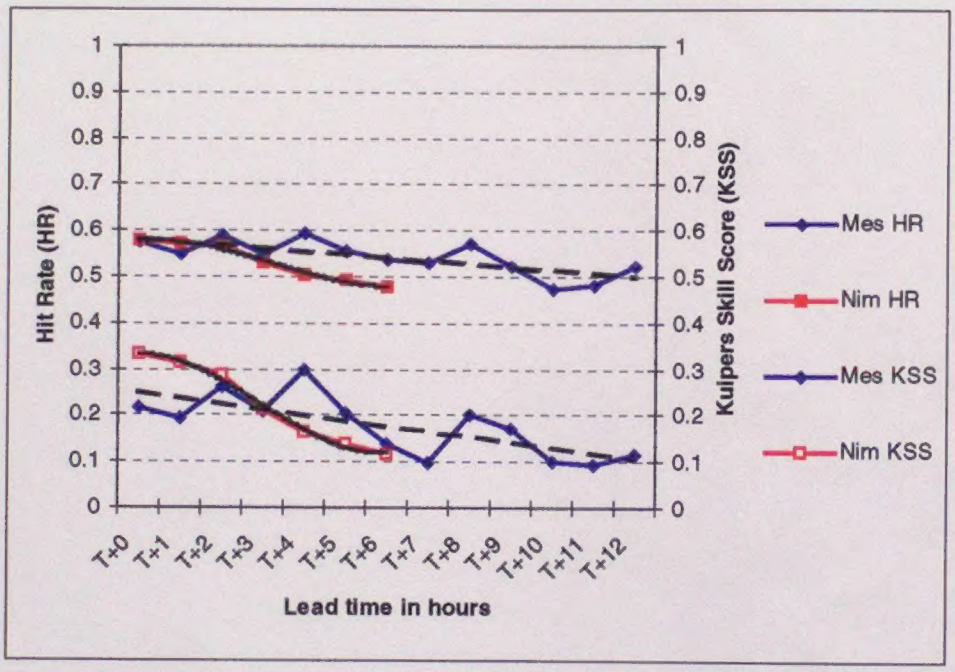
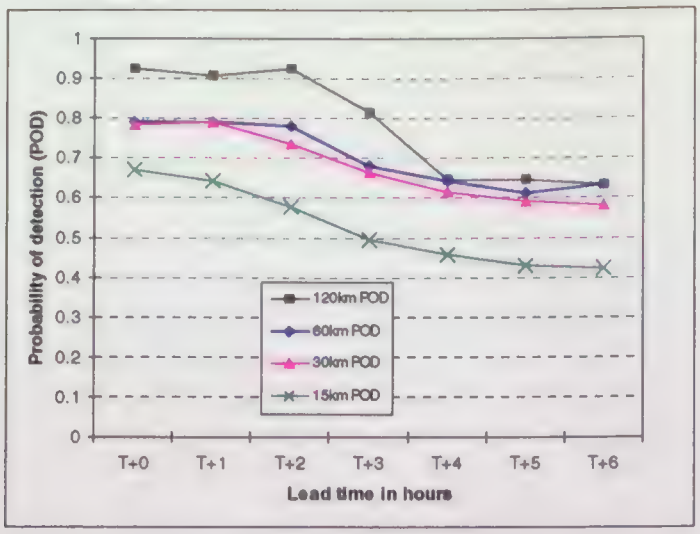


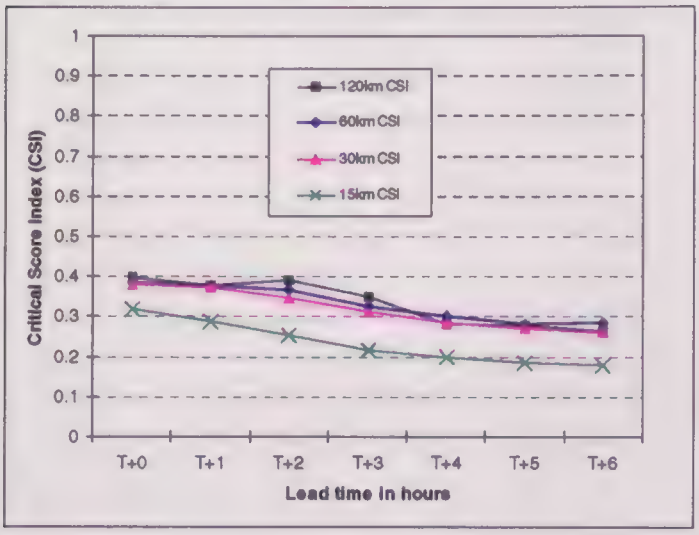
FIGURE 7: Hit Rate and Kuipers' Skill Score (KSS) for the disruption forecasts prepared from Nimrod and Mesoscale data (averaged over spatial resolution). Both forecast types decline in each measure over time, with the Nimrod forecasts performing better at the shorter lead times.

Thus it appears that between T+0 to T+3 the disruptions forecasts produced from the Nimrod data are of the higher quality. From T+4 to T+6 the results are closer, with the Mesoscale forecasts appearing to be better from T+4 to T+6. However, the Nimrod results are more reliable due to the lower level of scatter. Hence the details of the trade-off between spatial and lead-time accuracies were analysed for the Nimrod disruption forecasts only.

3.3 Spatial and lead-time accuracies for the Nimrod disruption forecasts

For Nimrod forecasts, figure 8 shows the POD, CSI and FAR for the 'Severe' category. The POD declines slowly with increasing lead time to T+2, steeply from T+2 to T+4, and levels off around T+5. The CSI behaves in a similar manner but at a lower level, while the FAR rises steadily with increasing lead time before reaching a plateau around T+5. The graphs also indicate that accuracy decreases as spatial resolution increases, with all four resolutions having similar patterns over lead time but at different levels of accuracy. The 15km resolution in particular has significantly lower level POD and CSI scores, and a significantly higher FAR.





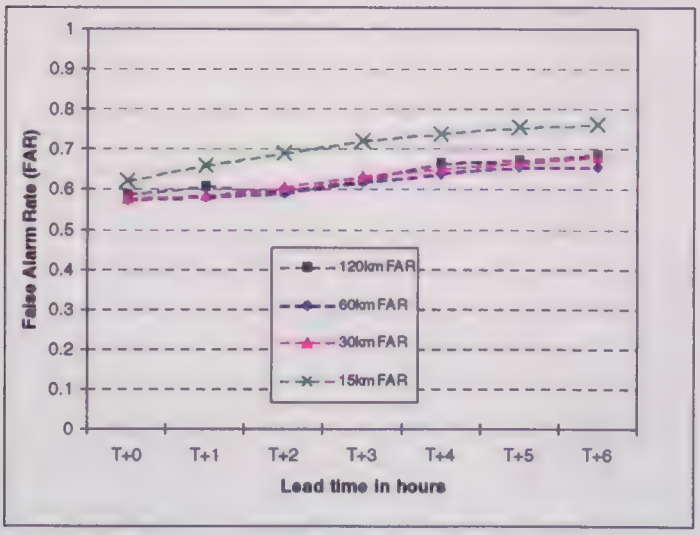


FIGURE 8: Comparing accuracy over lead time and resolution for 'Severe' air traffic disruption forecasts created using Nimrod data: POD (top), CSI (middle) and FAR (bottom). The POD and CSI decline as the lead time increases, particularly between T+2 and T+4, while the FAR rises steadily. The 15km spatial resolution has a significantly lower POD and CSI and a higher FAR.

The Nimrod Kuipers' Skill scores (the best overall measure of forecast skill which takes into account the overall quality of all four of the forecast levels of disruption) are plotted in figure 9. The four different spatial resolutions start with a KSS of between 0.3 and 0.4 at T+0 and decline to around 0.15 at T+6. The differences between the spatial resolutions are not so apparent in this graph because the greater resolutions benefit from a larger number of "forecast disruption level = 'None', actual disruption level = 'None'" situations. Even at T+6 the KSS remains around 0.15 for all resolutions, indicating the forecasts are still significantly better than random and may still be of some use in ATC management.

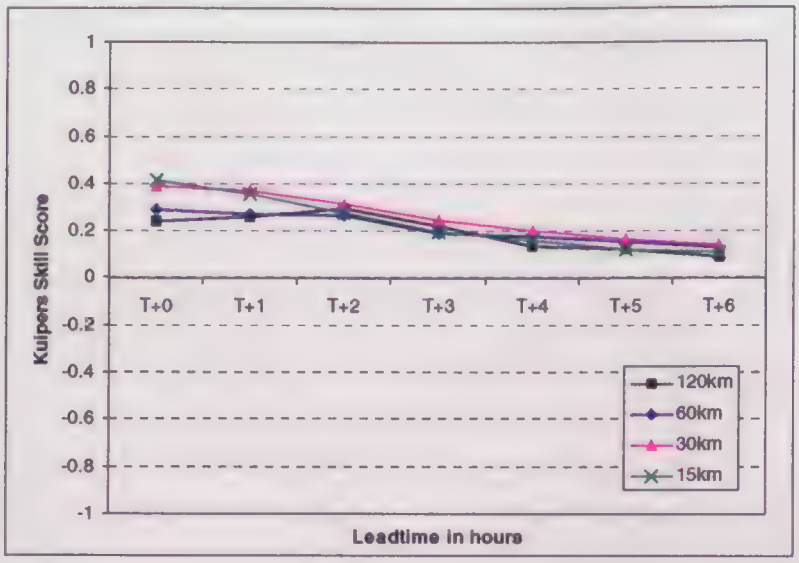


FIGURE 9: Kuipers' Skill Scores for the disruption forecasts produced using Nimrod data for each of the spatial resolutions considered. KSS declines steadily with lead time for all resolutions.

4 DISCUSSION

4.1 Trends in forecast quality

Lead time and spatial resolution
The drop off in forecast quality as lead time increases reflects the difficulty of forecasting small scale features such as thunderstorms over a long lead time. The reduction in accuracy with increased spatial resolution is also to be expected, as over smaller areas there is more scope for small errors to have a large impact.

Over-forecasting intensity
The forecasts are generally good at picking up the location and extent of the storms, but often overestimate their intensity (see figure 10 for an example). This tendency may cause the large bias value for the severe category forecasts, and may also explain why the method which includes coverage in its air traffic disruption forecasts, rather than just relying on intensity, performs better. Errors in forecasting the location of the strongest parts of the storm, or in picking up certain areas of activity, also restrict the quality.

Differences between thunderstorm events
The results analysed in the previous section were averaged over all the 10 events analysed. These average results hide many complexities in the quality of the disruption forecasts. For example, the forecast quality varies significantly between different events. One reason for this could be the synoptic conditions at the time, which varied significantly with event. For example, the thunderstorms over Northern France on 27 July 2001 were set off in the late afternoon after a very hot day associated with anticyclonic conditions, while those on the 14-16 June were associated with

a depression moving across the UK. The forecasts produced using Mesoscale and Nimrod data appear to perform significantly better in the former conditions (compare figure 11 with figure 10), possibly due to the problems associated with ruling out dynamic rain in the latter case, as discussed in section 4.2. Unfortunately the majority of the events for which data were available did not have the synoptic conditions of 27 July, and thus the overall results indicate a lower quality than might otherwise be the case. However, air traffic disruption due to thunderstorm activity will occur irrespective of the synoptic conditions and thus better ways of forecasting thunderstorm activity in all conditions must be found.

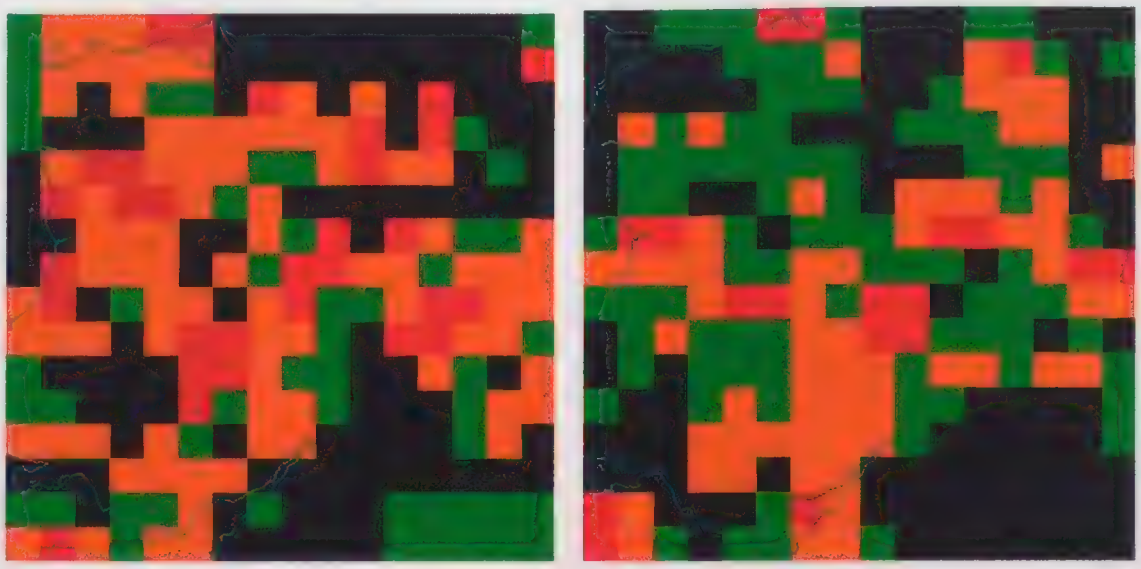


FIGURE 10: Replicated pilot displays for 1200 on 15 June 2001, created using Nimrod forecast data at T+1 (left), and radar data (right). The coverage of convective activity has been forecast adequately but the intensity has been overestimated for much of the analysis area.

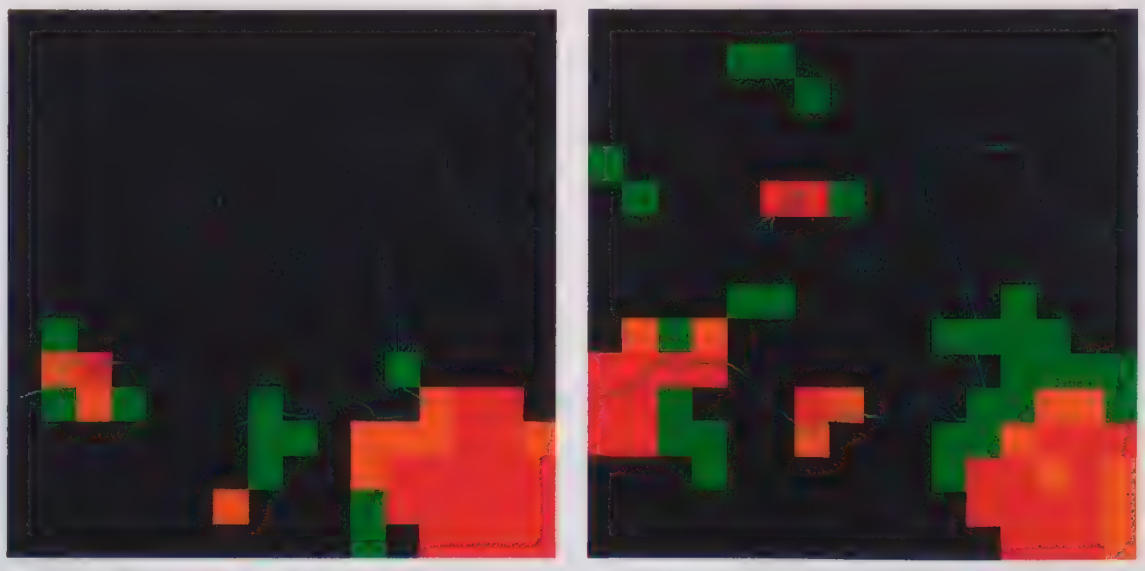


FIGURE 11: Replicated pilot displays for 1800 on 27 July 2001, created using Nimrod forecast data at T+0 (left), and radar data (right). Both the area and intensity of the storms have been successfully forecast.

Differences within thunderstorm events

The logarithmic profile of the idealised mature thunderstorm cell used to convert the rain rate at the ground to the maximum in the vertical makes the assumption that a mature cells is positioned in

each grid cell. This may not always be the case and as cells at different life stages have different profiles (Hand, 1996), this may limit the accuracy of the disruption forecasts. Indeed, in some cases the forecasts do appear to perform better in the middle of the event than at the start or the end.

Consequence of these trends for comparing Mesoscale and Nimrod disruption forecasts
Nimrod produces hourly output for each lead-time up to T+6. The Mesoscale model is only run every six hours, and thus for each hour only two Mesoscale lead times could be tested and the sample of analysis times for each lead time was much smaller. For example, the lead time of T+1 was able to be analysed 63 times using Nimrod data, compared to only 10 when using Mesoscale data. A consequence of this is that while the Nimrod results are independent of event and storm-stage, the Mesoscale results are dependent on which event/storm-stage combinations predominate at the hours when each lead time can be tested. As forecast accuracy appears to be dependent on both the event in question and the storm-stage, the sensitivity of the Mesoscale forecasts will have influenced their relative performance and may have increased the scatter in the quality of these disruption forecasts over the lead times considered.

4.2 Limitations of the project

There are several limitations of this project that will have influenced the accuracy of the results. The limitations apply equally to all the lead times analysed, including T+0, and help to explain why the T+0 forecasts do not perform as well as would normally be expected.

Radar data
The radar data processing has four limitations:

i. In some events the radar picture of maximum in-cloud rainfall does not appear to indicate severe storm activity. As the events were picked using lightning data, it is likely that the available radar beams have missed the maximum rainfall. This is a particular problem with radar sites that do not have suitable neighbouring sites, or that have low beam elevations, so that much of the radar beam data fell below the set cloud-base threshold of 3km. An example of this is the thunderstorm case on 7 July near the Hill of Dudwick radar in Northeast Scotland, for which replicated pilot displays are shown in figure 12. The grey area in the centre of the pictures is where no beams were available above the required height. Outside of this area only one or two beams were available and hence it is possible that the radar data missed the true maximum rainfall, and thus the forecast appeared to be of a poorer quality than may actually have been the case.

FIGURE 12: Replicated pilot displays for 1900 on 7 July 2001, created using Nimrod forecast data at T+1 (left), and radar data (right). The grey area in the centre of the pictures shows where no beams were available above the required height. Outside this area the limited radar beams available may have missed the true maximum rainfall, and falsely showed the forecasts to be of a poor quality.

ii. A radar beam is attenuated as it passes through rainfall or cloud. A correction for this was applied to the radar data used in this study, but could have been inaccurate in cases of severe attenuation. This means that the reflectivity of storms far away from the radar, such as those used from neighbouring sites, may have been incorrect.
iii. The radar data were not corrected for overestimates in radar reflectivity caused by melting, which may cause a 'bright band' in the radar data.
iv. An overestimate of reflectivity may also be caused by anomalous propagation. This occurs under certain weather conditions (such as temperature inversions), when the radar beam is bent downwards from its normal path, possibly intersecting the ground.

Together these problems mean that the data used for verifying the air disruption forecasts may itself not have given a true picture of the storms. However, in the absence of actual data of aircraft deviations, or of air traffic disruption due to thunderstorm activity, it is the best possible alternative.

Dynamic rain
Although steps were taken to eliminate dynamic rain from the forecast and verification data, these may not always have been effective. Thus the Nimrod forecasts, which used total rain rate masked with convective cloud cover, may compare better to the radar data which measured total rain rate, than the other two forecast types. Dynamic rain will be more problematic when predicting thunderstorms associated with an active depression, and may explain why the forecasts did not perform so well for these events.

5 CONCLUSIONS

This project has quantified the spatial and lead-time accuracies of forecasts of air traffic disruption due to thunderstorm activity for three different types of Met Office forecast. Forecast quality has been shown to be greatest for disruption forecasts prepared from Nimrod data, and following a method that uses both storm intensity and coverage to predict deviations. Forecast accuracy and skill decline with increasing lead time and spatial resolution up to T+5, and remain fairly constant at lead times greater than this. Forecast quality is not independent of event, and it is possible that the ability to forecast air traffic disruption depends on the synoptic conditions prevalent at the time of the thunderstorm event.

ATM authorities will need to interpret the results presented here in the context of their own operations. For example severe disruption may not be a problem if the density of traffic is low, or if the traffic is not going to impinge on other air space sectors. In addition, they will need to establish the costs and benefits involved in taking action based on forecasts of a given lead time and spatial resolution.

A follow-up project will extend this work to produce and verify air traffic disruption forecasts over summer 2002. Attempts will be made to improve the forecasting and verification methods, taking into account the limitations and forecast quality trends detailed in this report. Additional forecast data types will also be used, including the new GANDOLF forecasts, and an extrapolation forecast of multi-beam radar data.

6 REFERENCES

Golding, B.W., 1998: Nimrod: A system for generating automated very short range forecasts. *Meteorol Appl*, 5, 1-16.

Hand, W.H., Submitted: The UK Met Office Convection Diagnosis Scheme. *Meteorol Appl.*

Hand, W.H., 1996: An object-orientated technique for nowcasting heavy showers and thunderstorms. *Meteorol Appl*, 3, 31-41.

Hoad, D.J., 2000: Verification of short range thunderstorm forecasts using radar data to assess their benefit to the aviation community. *9th Conference on Aviation, Range, and Aerospace Meteorology*, 188-191.

Rankin, J., 2001: Personal communication.

Rhoda, D.A., Boorman, B.G., Bouchard, E.A., Isaminger, M.A., and Pawlak, M.L., 2000: Commercial aircraft encounters with thunderstorms in the Memphis terminal airspace. *9th Conference on Aviation, Range, and Aerospace Meteorology*, 37-42.

Rhoda, D.A. and Pawlak, M.L., 1999: The thunderstorm penetration/deviation decision in the terminal area. *8th Conference. on Aviation, Range, and Aerospace Meteorology*, 308-312.

Seymour, J., 2002: Personal communication.

Wilks, D.S., 1995: Statistical methods in the atmospheric sciences. London, Academic Press.

APPENDIX A

To follow are a list of the 10 thunderstorm cases analysed in this report.

Start date	Start time	End date	End time	Radar site that the analysis area is centred upon
14/06/2001	1400	14/06/2001	2000	Hameldon Hill
15/06/2001	1000	15/06/2001	1300	Clee Hill
15/06/2001	1400	15/06/2001	2000	Ingham
16/06/2001	0900	16/06/2001	1600	Chenies
03/07/2001	1900	04/07/2001	0000	Hameldon Hill
07/07/2001	0900	07/07/2001	1200	Hill of Dudwick
07/07/2001	1300	07/07/2001	2100	Hill of Dudwick
07/07/2001	1300	07/07/2001	2100	Chenies
07/07/2001	2200	08/07/2001	0200	Hill of Dudwick
27/07/2001	1600	27/07/2001	2000	La Moye

APPENDIX B

Below is an example of a 4 x 4 contingency table. Each box gives the number of times a particular combination of observed and forecast disruption level occurs. For example q represents the number of times that the air traffic disruption due to thunderstorm activity was forecast to be severe and observed to be severe.

		Observed				
		None	Minor	Major	Severe	
Forecast	None	a	b	c	d	ε=a+b+c+d
	Minor	e	f	g	h	ζ=e+f+g+h
	Major	i	j	k	l	η=i+j+k+l
	Severe	m	o	p	q	θ=m+o+p+q
		α=a+e+i+m	β=b+f+j+o	γ=c+g+k+p	δ=d+h+l+q	n=a+b+c+d+e+f+g+h+i +j+k+l+m+o+p+q

Forecast quality measures can be calculated from the above table using the following equations.

$$Hit\ Rate = \frac{a+f+k+q}{n}$$

$$Bias\ (none) = \frac{\varepsilon}{\alpha} \qquad Bias\ (minor) = \frac{\zeta}{\beta} \qquad Bias\ (major) = \frac{\eta}{\gamma} \qquad Bias\ (severe) = \frac{\theta}{\delta}$$

$$POD\ (none) = \frac{a}{\alpha} \qquad POD\ (minor) = \frac{f}{\beta} \qquad POD\ (major) = \frac{k}{\gamma} \qquad POD\ (severe) = \frac{q}{\delta}$$

$$CSI\ (none) = \frac{a}{\alpha+b+c+d} \quad CSI\ (minor) = \frac{f}{\beta+e+g+h} \quad CSI\ (major) = \frac{k}{\gamma+i+j+l} \quad CSI\ (severe) = \frac{q}{\delta+m+o+p}$$

$$FAR\ (none) = \frac{b+c+d}{\varepsilon} \qquad FAR\ (minor) = \frac{e+g+h}{\zeta} \qquad FAR\ (major) = \frac{i+j+l}{\eta} \qquad FAR\ (severe) = \frac{m+o+p}{\theta}$$

$$KSS = \frac{\left[\left(\frac{a}{n}\right)+\left(\frac{f}{n}\right)+\left(\frac{k}{n}\right)+\left(\frac{q}{n}\right)\right]-\left[\left(\frac{\varepsilon}{n}\right)*\left(\frac{\alpha}{n}\right)+\left(\frac{\zeta}{n}\right)*\left(\frac{\beta}{n}\right)+\left(\frac{\eta}{n}\right)*\left(\frac{\gamma}{n}\right)+\left(\frac{\theta}{n}\right)*\left(\frac{\delta}{n}\right)\right]}{1-\left[\left(\frac{\alpha}{n}\right)^2+\left(\frac{\beta}{n}\right)^2+\left(\frac{\gamma}{n}\right)^2+\left(\frac{\delta}{n}\right)^2\right]}$$